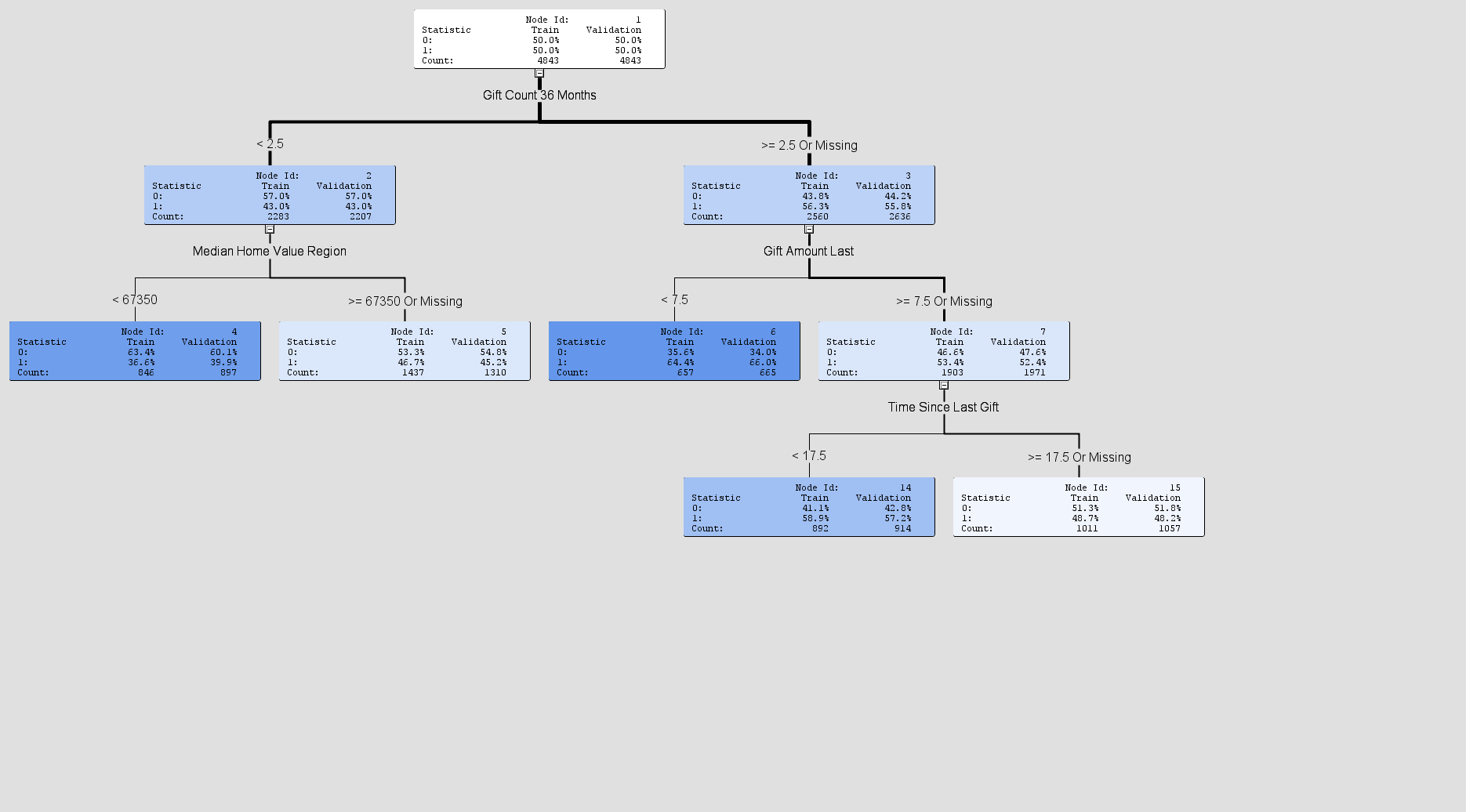
**In-Class Exercise: Understanding Decision Tree Output**

The following is the output from a decision tree analysis in SAS Enterprise Miner. It uses the same data set (a charity donation database) you’ll be using in your in-class exercise, but I’ve changed the model parameters so we get some different results.

****

**Labels:**

**Gift Count 36 Months** – Number of donations the person made in the last 36 months

**Gift Amount Last** – Value of the last donation the person made (in $ thousands)

**Time Since Last Gift** – Number of months since the person’s last donation

**Median Home Value Region** – The median home value in the donor’s geographic region

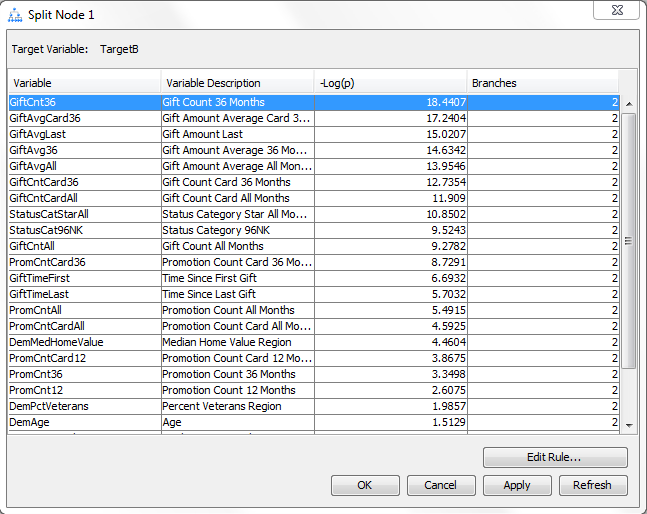
**Outcome statistic (look at the validation set results when evaluating)**:

0 = person doesn’t donate an additional gift within the next 3 months

1 = person does donate an additional gift within the next 3 months

Answer the following questions about this tree on the next page:

1. What is the probability that the following people will donate an additional gift in the next 3 months?  
   1. Someone who has made two donations in the last 36 months and the median home value in their region is $85,000.
   2. Someone who has made five donations in the last 36 months and their last gift amount was $3,000.
   3. Someone who has made three donations in the last 36 months, their last gift amount was $10,000, and it’s been 12 months since their last donation.
   4. Someone who has made three donations in the last 36 months, their last gift amount was $10,000, and it’s been 24 months since their last donation.
2. What does it mean that median home value isn’t a split variable when more than three donations were made in the last 36 months?
3. Check out this screen shot from SAS Enterprise Miner. This is the list of logworth values used to choose the first split variable for the data.  
   1. According to this list, which variable is the best one to choose for that first split? Why?
   2. How does this correspond to the final decision tree on the previous page?



1. Remember the logworth value is the log of the p-value derived from the Chi-Squared statistic. So the logworth value is directly related to a Chi-Squared calculation comparing observed outcomes versus expected outcomes.  
     
   With this in mind, consider these two hypothetical scenarios for a different set of donor data:  
     
   Scenario 1: Gender as a split variable

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Observed | | |  |  | Expected | | |
|  | Male | Female |  |  |  | Male | Female |  |
| Gift | 375 | 125 | 500 |  | Gift | 400 | 100 | 500 |
| No Gift | 425 | 75 | 500 |  | No Gift | 400 | 100 | 500 |
|  | 800 | 200 | 1000 |  |  | 800 | 200 | 1000 |

Scenario 2: Age as a split variable

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Observed | | |  |  | Expected | | |
|  | <40 | >=40 |  |  |  | <40 | >=40 |  |
| Gift | 405 | 95 | 500 |  | Gift | 400 | 100 | 500 |
| No Gift | 395 | 105 | 500 |  | No Gift | 400 | 100 | 500 |
|  | 800 | 200 | 1000 |  |  | 800 | 200 | 1000 |

a) What is the Chi-Squared value for Scenario 1 and Scenario 2?  
  
Scenario 1 (Gender): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Scenario 2 (Age): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
  
  
b) Which split variable, Gender or Age, is a more powerful differentiator whether someone will donate a gift?  
  
Answer: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
  
  
c) Briefly explain why:  
  
Answer: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
  
  
d) Which split variable, Gender or Age, will have the higher logworth value?  
  
Answer: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_